FOR THE ABANDONMENT OF SYMMETRY IN THE THEORY OF COOPERATIVE GAMES

T. C. Schelling

P-1386

29 May 1958

The RAND Corporation
700 MAIN ST. • SANTA MONICA • CALIFORNIA

CLEARINGHOUSE FOR FEDERAL SCIENTIFIC AND TECHNICAL INFORMATION CFSTI
DOCUMENT MANAGEMENT BRANCH 410.11

LIMITATIONS IN REPRODUCTION QUALITY

ACCESSION # AD 606728

☒ 1. WE REGRET THAT LEGIBILITY OF THIS DOCUMENT IS IN PART UNSATISFACTORY. REPRODUCTION HAS BEEN MADE FROM BEST AVAILABLE COPY.

☐ 2. A PORTION OF THE ORIGINAL DOCUMENT CONTAINS FINE DETAIL WHICH MAY MAKE READING OF PHOTOCOPY DIFFICULT.

☐ 3. THE ORIGINAL DOCUMENT CONTAINS COLOR, BUT DISTRIBUTION COPIES ARE AVAILABLE IN BLACK-AND-WHITE REPRODUCTION ONLY.

☐ 4. THE INITIAL DISTRIBUTION COPIES CONTAIN COLOR WHICH WILL BE SHOWN IN BLACK-AND-WHITE WHEN IT IS NECESSARY TO REPRINT.

☐ 5. LIMITED SUPPLY ON HAND: WHEN EXHAUSTED, DOCUMENT WILL BE AVAILABLE IN MICROFICHE ONLY.

☐ 6. LIMITED SUPPLY ON HAND: WHEN EXHAUSTED DOCUMENT WILL NOT BE AVAILABLE

☐ 7. DOCUMENT IS AVAILABLE IN MICROFICHE ONLY.

☐ 8. DOCUMENT AVAILABLE ON LOAN FROM CFSTI (TT DOCUMENTS ONLY).

☐ 9.

NBS 9 64 PROCESSOR: PL

FOR THE ABANDONMENT OF SYMMETRY IN THE THEORY OF COOPERATIVE GAMES

T. C. Schelling

The first part of this paper argues that the pure, "moveless" bargaining game analyzed by Nash, Harsanyi, Luce and Raiffa, and others, may not exist or, if it does, is of a different character from what has been generally supposed; the point of departure for this argument is the operational meaning of agreement, a concept that is almost invariably left undefined. The second part of the paper argues that symmetry in the solution of bargaining games cannot be supported on the notion of "rational expectations"; the point of departure for this argument is the operational identification of irrational expectations.

Part I

The logical structure (move structure) of a zero-sum game is completely defined by the game's payoff matrix or function; and, if one is attracted to the minimax solution, the payoff matrix contains everything significant in the game. For the tacit (non-cooperative) non-zero-sum game, though not all of the essentials of the game are necessarily contained in the payoff matrix, the entire move structure is. There are no "moves" in this game, except the unilateral selection of strategies, and the timing of such selection is immaterial since there is no outcome until both players have selected. But the non-tacit (cooperative) non-zero-sum game is not defined by its

payoff matrix; the operations by which choices are made must still be specified. Commonly these operations are sketched in by reference to the notion of "binding agreements," together with the notion of free communication in the process of reaching agreement. Thus to say that two players may divide $100 as soon as they can agree on how to divide it, and that they may discuss the matter fully with each other, is generally considered sufficient to define a game.*

A game of this sort is symmetrical in its "move" structure, even though it may be asymmetrical in the configuration of payoffs. The two players have identical privileges of communication, of refusing offers, and of reaching agreement. If instead of dividing $100 the players are to agree on values X and Y contained within a boundary, the payoff function may not be symmetrical but the move structure is. Harsanyi, to emphasize this, has even added explicitly the postulate of symmetrical moves: "The bargaining parties follow identical (symmetric) rules of behaviour (whether because they follow the same principles of rational behaviour or because they are subject to the same psychological laws)."**

---

*Luce and Raiffa, in effect, define cooperative two-person games by reference to a payoff matrix and the following three stipulations:

(i) All preplay messages formulated by one player are transmitted without distortion to the other player.

(ii) All agreements are binding, and they are enforceable by the rules of the game.

(iii) A player's evaluations of the outcomes of the game are not disturbed by these preplay negotiations.

Games and Decisions, p. 114.

**John Harsanyi, "Approaches to the Bargaining Problem Before and After the Theory of Games...," Econometrica, Vol. 24, April 1956, p. 149.

FOR THE ABANDONMENT OF SYMMETRY IN THE THEORY OF COOPERATIVE GAMES

T. C. Schelling

The first part of this paper argues that the pure, "moveless" bargaining game analyzed by Nash, Harsanyi, Luce and Raiffa, and others, may not exist or, if it does, is of a different character from what has been generally supposed; the point of departure for this argument is the operational meaning of agreement, a concept that is almost invariably left undefined. The second part of the paper argues that symmetry in the solution of bargaining games cannot be supported on the notion of "rational expectations"; the point of departure for this argument is the operational identification of irrational expectations.

Part I

The logical structure (move structure) of a zero-sum game is completely defined by the game's payoff matrix or function; and, if one is attracted to the minimax solution, the payoff matrix contains everything significant in the game. For the tacit (non-cooperative) non-zero-sum game, though not all of the essentials of the game are necessarily contained in the payoff matrix, the entire move structure is. There are no "moves" in this game, except the unilateral selection of strategies, and the timing of such selection is immaterial since there is no outcome until both players have selected. But the non-tacit (cooperative) non-zero-sum game is not defined by its

What I want to do now is to look at this notion of "agreement" on the assumption of perfect symmetry in the move structure of the game, paying close attention to the "legal details" of the bargaining process. We must also look at the meaning of "nonagreement." Since any well-defined game must have some rule for its own termination, let us look at the rules for termination first.

If we are to avoid adding a whole new dimension to our payoff matrix, in the form of discount rates, we must suppose that the game is terminated soon enough so that nothing like the interest rate enters the picture. We do not want to have to consider the time at which agreement is reached, in addition to the agreement itself. This is not only a matter of convenience; the game ceases to be "moveless" except in very special cases, unless we make this stipulation. For if the players' time preferences take any shape except that of a continuously uniform discount rate, the game itself changes with the passage of time and a player can, in effect, change the game itself by failing to reach agreement. The notion of a continuously uniform discount rate is far too special and unrealistic to use as a basic postulate; so we must assume that the game is somehow gotten over with.

Perhaps the simplest way to terminate the game is to have a bell ring at a specified time known in advance. There are other ways, such as having the referee roll dice every few minutes, calling off the game whenever he rolls boxcars. (We might have the game terminate after a specified number of offers have been refused, but this would change the character of the game by making certain kinds of communication "real moves" that leave the game different from what it was before, and

perforce lead us into such tactics as the exhaustion of offers.)

For simplicity, suppose that the game will be terminated at a specified time, known in advance to the players, and for convenience of discussion let us call that final moment "midnight." If agreement exists when the midnight bell rings, the players divide the gains in the way they have agreed; if no agreement exists, the players receive nothing.

Next, what do we mean by "agreement?" For simplicity, suppose that each player keeps (or may keep) his current "Official" offer recorded in some manner that will be visible to the referee when the bell rings. Perhaps he keeps it written on a blackboard that the other player can see; perhaps he keeps it in a sealed envelope that is surrendered to the referee when the bell rings; perhaps he keeps it punched into a private keyboard that records his current offer in the referee's room. When the bell rings, the blackboard is photographed, the envelope surrendered, or the keyboard locked, so that the referee only needs to inspect the two "current" offers as they exist at midnight to see whether they are compatible or not. If they are compatible, the gains are divided in accordance with the "agreement"; if the two players have jointly claimed more than is available, "disagreement" exists and the players get nothing. (Defer, for a moment, ruling on what happens if the two players together have claimed less than the total available, whether they get as much as they have claimed or get nothing for lack of proper agreement. And, in what follows, it will not matter whether an exhaustive agreement reached before midnight -- i.e., compatibility of the current offers occurring before midnight --

terminates the game.)

There are other ways of defining "agreement" in terms of the operations by which it is reached or recorded; but if we adhere to the notion of a symmetrical move structure they will generally, I think, have the property that I am trying to single out for attention. That property is this. There must be some minimum length of time that it takes a player to make, or to change, his current offer. (For simplicity again, let us suppose that the same operation either makes an offer or changes it, so that we may always assume that a "current offer" exists.) There must then be some critical moment in time, a finite period before the midnight bell rings, that is the last moment at which a player can begin the operations that record his final offer. That is, there is some last moment before the bell rings, beyond which it is too late to change one's existing offer. Under the rules of the game and the rationality postulate both players know this. And by the rule of symmetry this moment must be the same for both players.

From this follows the significant feature. The last offer that it is physically possible for a player to make is one that he necessarily makes without knowing what the other player's final offer is going to be; and the last offer that a player can make is one that the other player cannot possibly respond to in the course of the game. Prior to that penultimate moment, no offer has any finality; and at that last moment players either change or do not change their current offers, and whatever they do is done in complete ignorance of what each other is doing, and is final.*

*Incidentally, the argument is unaffected by supposing that a player can change his offer "instantaneously" as long as we keep the symmetrical rule that both can do it "equally instantaneously" as the final bell rings.

This must be true. If either got a glimpse of the other's final offer in time to do anything about it, or if either could give the other a glimpse of his own final offer in time for the other to respond, it was not -- and is known to be not -- a final offer.*

But now we have reached an important conclusion about the perfectly move-symmetrical bargaining game. It is that it necessarily gives way, at some definite penultimate moment, to a tacit (non-cooperative) bargaining game. And each player knows this. The most informative way to characterize the game, then, is not that the players must reach overt agreement by the time the final bell rings or forego the rewards altogether. It is that they must reach overt agreement by a particular (and well identified) penultimate moment -- when the "warning bell" rings -- or else play the tacit game with the same payoff matrix.

Each player must be assumed to know this and may, if he wishes, by simply avoiding overt agreement, elect to play the tacit game instead. So if we assume for the moment that the tacit game has a clearly recognized solution, and that the solution is efficient, each player has a pure minimax behavior strategy during the earlier stage. Either can enforce this tacit solution by abstaining from agreement until the warning-bell rings; neither can achieve anything better from a rational opponent by verbal bargaining.

From this it follows that the solution of the cooperative game must

---

*There is a mechanical assumption here that in the process of making a new offer one can stop and start over. The case is slightly more complicated if an offer started one and one-half minutes before midnight is necessarily the last offer because the process cannot be started again until a minute has passed and by then the critical point has been passed. This case will be looked at again below.

be identical with that of the corresponding tacit game, if the latter has a predictable and efficient solution. It must, because the tacit game comes as an inevitable, mechanical sequel to the cooperative game. At this point it looks as though the cooperative feature of the game is irrelevant; the players really need not show up until 11:59, in fact they do not need to show up at all. The "preplay communication" and ability to reach "binding agreements," which were intended to characterize the game, prove to be irrelevant; the "cooperative game" as a distinct game from the tacit game does not exist.*

But this conclusion is unwarranted. First, a tacit game may not have a confidently predicted efficient solution. More than that, certain details of the cooperative game that might have seemed to be innocuous from the point of view of explicit negotiation may affect the character of the tacit game; similarly, preplay communication that has no binding effect on the players themselves may also affect the character of the tacit game. Just for example, consider the following variant to the cooperative game.

Instead of saying that the players may divide a set of rewards if

---

*In his 1953 article, "Two-person Cooperative Games," J. F. Nash presents a model that is explicitly tacit in its final stage. The model's relation to the cooperative game was heuristic: it was to help to discover what might constitute "rational expectations" (and hence the indicated rational outcome) in the corresponding cooperative game. The argument of the present paper is that the relation is likely to be mechanical rather than intellectual if a symmetrical move-structure is strictly adhered to, and that with strict symmetry it is difficult, perhaps impossible, to define the corresponding non-tacit game that was the ultimate subject of study. (Nash, Econometrica, Vol. 21, pp. 126-140.)

they can reach agreement on an exhaustive division, let us say that the players may divide a set of rewards to the extent that they have reached agreement on a division; they may divide such portion of the available rewards as they have already reached agreement on by the time the bell rings. If, for example, there are one hundred indivisible objects and they have reached agreement on how to divide eighty of them when the bell rings, the twenty items in dispute revert to the house while the eighty on which agreement was reached will be divided in accordance with the agreement.*

Now, in the cooperative case, if we had already concluded there was an efficient solution to this game -- i.e., that the players would in fact reach an exhaustive agreement -- we should probably have considered this reformulation of the problem inconsequential. It only says, in effect, that bargaining should take the form of each player's writing down the totality of his claim and that concessions shall take the form of each player's deleting items from his list of claims, with full agreement's being reached when no more items are in conflict on the lists of claims. But when we look at the tacit case, the game is drastically altered by this reformulation. The tacit game now has a perverse incentive structure. There is no rational reason for either player to demand less than the whole of the available reward; each knows this and knows that the other knows it. There is no incentive to reduce

---

*In the case of a single divisible object like money, the corresponding rule might be that they divide the money in accordance with their offers after the house has removed the "overlap." Each player obtains as much as the other implicitly accords him, if one is demanding 65 percent of the money at the end of the game, and the other 55 percent, the second has been accorded 35 percent and the first 45 percent, these amounts are outside the range of dispute and constitute the "agreement."

one's claim because any residual dispute costs the player no more than he would lose if he reduced his claim to eliminate the dispute. The single equilibrium point yields zero for both players. Thus the variant game, which seemed to differ inconsequentially, is drastically different from the original game; but it does not appear so until we have identified the terminal tacit game as a dominating influence.*

To take another example, suppose there are 100 individual objects to be divided and that, although they are fungible as far as value is concerned, the agreement must specify precisely which individual items go to which individual players. If the rules require that full and exhaustive agreement be reached, then in the tacit game the players are dependent on their ability not only to divide the total value of the objects in coordinated fashion but to sort out the 100 individual objects into two piles in identical fashion. If, then, one of the players has demanded specific items worth 60 percent of the total and the other player has refused, the former has an advantage in the tacit game. The only extant proposal for dividing the 100 objects is the one

---

* It might seem that we can draw a by-product from the analysis here, namely, the observation that in order to set up a "truly" cooperative (non-tacit) game, the legal definition of agreement must be such as to make the ultimate tacit game perverse, so that the players must reach binding agreement before the warning bell or suffer complete loss. But there is still a problem. Assuming that the players themselves can define "agreement" for purposes of agreement prior to the final ball, and that the perverse rule only governs the definition of "agreement" at termination if no prior agreement exists, we must now provide (or assume the players to provide) an operational definition of agreement. If it is like our earlier definition, all they accomplish is to make the perverse cooperative game into a benign one, one minute shorter, which is equivalent to a tacit game two minutes shorter than the original, and the solution to the tacit game governs if it is confidently foreseen and efficient.

player's specification of 80 that would satisfy him; the chances of their concerting identically on any other division of the 100 objects, equal or unequal between them, may be so small that they are forced for the sake of agreement into accepting the only extant proposal in spite of its bias. Thus preplay communication has tactical significance in that it can affect the means of coordination once the tacit stage of the game has been reached.

If now, in considering the tactical implications of this last point, we insist on a rule of symmetrical behavior, we must conclude that if either player opened his mouth to drown out what the other was about to say, he would always find the other player also with his mouth open, both knowing that if either spoke the other would be found to be speaking, neither able to hear the other, and so on. In other words, the assumption of complete symmetry of behavior as a recognized foregone conclusion seems to preclude the very kind of action that might have seemed to enrich the game at the stage of preplay communication.

But by now we have certainly pressed the perfect move-symmetrical game as far as is worthwhile.* We could go on to analyze this game in

---

*One detail may be worth pursuing, in line with an earlier footnote. Suppose that it takes one minute to make or change an offer and (in contrast to the earlier version) that the process of recording a new offer, once started, cannot be stopped before it is completed. Under this procedure, any offer initiated during the next to last minute of the game is one's final offer. If this final offer cannot be communicated to the other player before the expiration of the minute, the game is essentially the same as before. "simultaneous" now means within a minute of each other for practical purposes, and again neither can see the other's final offer as he initiates his own, no matter what time during the final minute the offers are initiated. But suppose one punches his offer into a visible board which remains locked for one minute while the offer is recorded, so that the other player can see one's offer in a few seconds although one cannot initiate a change until the minute's delay is

more detail, considering such things as alternative ways of terminating the game or of defining "agreement," etc.; it seems more worthwhile, however, to raise at this point the question of whether the perfectly "moveless" or "move-symmetrical" game is a profitable one to study. Is the nondiscriminatory, move-symmetrical game a "general" game, one that gets away from "special cases." Or is it a special, limiting case in which the most interesting aspects of the cooperative game have vanished?

It should be emphasized that the fruitful alternative to symmetry is not the assumption of asymmetry, but just nonsymmetry, admitting both symmetry and asymmetry as possibilities without being committed to either as a foregone conclusion.

An illustration may help. Suppose we were to analyze the game in which there is $100 at the end of the road for the player who can get there first. This game of skill is not hard to analyze: the money goes to the fastest, barring accidents and random elements. We can predict rational behavior (running) and the outcome (money to the fastest). Ties

---

up. In this case, if the two offers during that final minute are not simultaneous, the player who moves second makes his final offer in full knowledge of the other's, and since his only chance of winning anything is to accept it, he must accept whatever the other has offered. Thus "second move" loses if the first mover knows that the other is waiting. We now have a game that can be characterized as follows: the players dally around for 23 hours 59 minutes and then play a game lasting one minute, this game allowing each player one and only one offer which he can make at any time during the minute. This game offers, in effect, three strategies to a player, namely, (1) assume the other will wait, and demand 99 percent; (2) assume both will make simultaneous offers, and demand whatever is indicated by the "tacit" game; (3) wait. If both wait, the game is still to be played. If there is a finite number of potential "waits," we have strategies of wait-once-then-demand-99-percent, wait-once-then-demand-tacit-solution, wait-twice-demand-99-percent, wait-twice-demand-tacit-solution, etc. This game (the "tacit supergame" consisting of all strategies for playing the one-minute game) is then the game; and it has, if we wish to accept it, its own "solution in the strict sense" which consists of all strategies (all lengths of waits) that end in demands that correspond to the solution of the tacit game.

will occasionally occur; but they will occur at the end of a race and will not be taken for granted at the outset. We need an auxiliary rule to cover ties, but it would not dominate either the game or the analysis.

Consider the same game played in a population in which everybody can run exactly as fast as anybody else, and everybody knows it. Now what happens? Every race ends in a tie, so the auxiliary rules is all that matters. But since a tie is foregone conclusion, why would they bother to run?

The perfectly moved-symmetrical cooperative game seems a little like that foot race. Barga'ning in the one case is as unavailing as leg-work in the other; every player knows in advance that all moves and tactics are foredoomed to neutralization by the symmetrical potentialities available to his opponent. The interesting elements that we might inject in the bargaining game are meaningless if perfect symmetry, and its acceptance as inevitable by both players, are imposed on the game by its definition.

What should we add to the game to enrich it if the assumption of symmetry is dropped? There are many "moves" that are often available, but not necessarily equally available to both players, in actual game situation. "Moves" would include commitments, threats, promises; tampering with the communication system; invocation of penalties on promises, commitments, and threats; conveyance of true information, self-identification, and the injection of contextual detail that may constrain expectations, particularly when communication is incomplete.

To illustrate, suppose in the earlier cooperative game there is a turnstile that permits a player to leave but not to return; his current offer as he goes through the turnstile remains on the books until the bell rings. Now we have a means by which a player can make a "final"

offer, a "commitment;" whoever can record an offer favorable to himself and known to the other, and leave the room, has the winning tactic. Of course it may win for either of them; but this may mean that we end up with something like a foot race, and the one closest to the turnstile wins. By analyzing the tactic, and its institutional or physical arrangements, we may determine who can make first use of it.

We have not, it should be noted, converted the game of strategy into a game of skill by letting them race for the turnstile. It remains true that one wins when he gets to the turnstile first only through the other's cooperation, only by constraining the other player's choice of strategy. He does not win legally or physically by going through the turnstile; he wins strategically. He makes the other player choose in his favor. It is a tactic in a game of strategy, even though the use of it may depend on skill or locational advantage.

We can even put a certain kind of symmetry into the game now, without destroying it; we can flip a coin to see who is nearest the turnstile when the game begins, or let the players be similarly located and similar of speed but with random elements to determine who gets to the turnstile first. Though the game is now non-discriminatory, the outcome would still be asymmetrical because each player has an incentive to run to the turnstile, leaving behind a standing offer in his own favor.

We can include some risk of "tie," especially if there are two turnstiles and the players might go to them simultaneously. This constitutes "symmetry" as an interesting possibility, but not as a foregone conclusion; stalemate and the anticipation of it become interesting possibilities if the actions and information structure are in fact

conducive to ties. But with non-symmetry as our philosophy, we do not need to be obsessed with the possibility of ties.

Again, if one player can make an offer and destroy communication, he may thereby win the ensuing tacit game by having provided the only extant offer that both players can converge on when they badly need to concert their choices later during the final tacit stage. To be sure, we can consider what happens when identical capacities for destruction of communication are present, and both players must recognize that they may simultaneously destroy communication without getting messages across; but this interesting case seems to be a special one, not the general case.

In summary, the perfectly "moveless" or "move-symmetrical" cooperative game is not a very fruitful one to study, but rather a limiting case that may degenerate into an ordinary tacit game. The cooperative game is rich and meaningful only when "moves" are admitted; and even much of the significance of the moves will vanish if complete symmetry in the availability to the players is stamped into the definition of the game. It is the moves that are interesting, not the game without moves; and it is the potential asymmetry of the moves that makes them significant.

Part II

Symmetry is not only commonly imposed on the move-structure of games but adduced as a plausible characteristic of the solution of the game or of the rational behavior with which the solution must be consistent. Nash's theory of the two person cooperative game explicitly postulates symmetry.

as does Harsanyi's.* The symmetry postulate is certainly expedient; it often permits one to find a "solution" to a game and to stay -- if he wishes to -- within the realm of mathematics. There are few similarly potent concepts that compete with it as bases for solving a game. But the justification for the symmetry postulate has not been just that it leads to nice results; it has been justified on grounds that the contradiction of symmetry would tend to contradict the rationality of the two players. This is the underpinning that I want to attack.

What I am going to argue is that though symmetry is consistent with the rationality of the players, it is not possible to demonstrate that asymmetry is inconsistent with their rationality. I shall argue that the identification of symmetry with rationality constitutes "implicit theorizing" with non-operational concepts. I then want to offer what I think is an argument in favor of symmetrical solutions, which tends to make it but one of many potential influences on the outcome with no prima facie claim to pre-eminence.

Explicit statements of the relation between symmetry and rationality have been given by John Harsanyi. He says, "The bargaining problem has an obvious determinate solution in at least one special case: viz., in situations that are completely symmetric with respect to the two bargaining parties. In this case it is natural to assume that the two parties will tend to share the net gain equally since neither would be prepared to grant

---

* It is not always clear whether symmetrical behavior, as prescribed for example in the earlier quotation from Harsanyi, is to be considered a rule of the game under analysis or a behavior postulate separate from the definition of the game; but for the purpose of the present argument it will not matter.

the other better terms than the latter would grant him."* In a later paper he refers to the symmetry axiom as the "fundamental postulate" and says, "Intuitively the assumption underlying this axiom is that a rational bargainer will not expect a rational opponent to grant him larger concessions than he would make himself under similar conditions."**

Now this "intuitive" formulation involves two postulates. First, that one bargainer will not concede more than he would expect to get if he himself were in the other position. Second, that the only basis for his expectation of what he would concede if he were in the other position is his perception of symmetry.

The intuitive formulation, or even a careful formulation in psychological terms, of what it is that a rational player "expects" in relation to another rational player, poses a problem in sheer scientific description. Both

---

* John C. Harsanyi, "Approaches to the Bargaining Problem Before and After the Theory of Games: A Critical Discussion of Zeuthen's, Hicks', and Nash's Theories," Econometrica, Vol. 24, No. 2, April, 1956, p. 147. He goes on to say, "For instance, everybody will expect that two duopolists with the same cost functions, size, market conditions, capital resources, personalities, etc., will reach an agreement giving equal profits to each of them."

** The full quotation deserves to be given: "What the Zeuthen-Nash theory of bargaining essentially proposes to do is to specify what are the expectations that two rational bargainers can consistently entertain as to each other's bargaining strategies if they know each other's utility functions. The fundamental postulate of the theory is a symmetry axiom, which states that the functions defining the two parties' optimal strategies in terms of the data (or, equivalently, the functions defining the two parties' final payoffs) have the same mathematical form, except that, of course, the variables associated with the two parties have to be interchanged. Intuitively the assumption underlying this axiom is that a rational bargainer will not expect a rational opponent to grant him larger concessions than he would make himself under similar conditions."

Harsanyi, "Bargaining in Ignorance of the Opponent's Utility Function," Cowles Foundation Discussion Paper No. 46, December 11, 1957.

players, being rational, must recognize that the only kind of "rational" expectation they can have is a fully shared expectation of an outcome. It is probably not quite accurate -- as a description of the psychological phenomenon -- to say that one expects the second to concede something or to accept something; the second's readiness to concede or to accept is only an expression of what he expects the first to accept or to concede, which in turn is what he expects the first to expect the second to expect the first to expect, and so on. To avoid an "ad infinitum" in the descriptive process, we have to say that both sense a shared expectation of an outcome; one's "expectation" is a belief that both identify the same outcome as being indicated by the situation, hence as virtually inevitable. Both players, in effect, accept a common authority -- the power of the game to dictate its own solution through their intellectual capacity to perceive it -- and what they "expect" is that they both perceive the same solution.*

* Viewed in this way, the intellectual process of arriving at "rational expectations" in the full-communication bargaining game is virtually identical with the intellectual process of arriving at a coordinated choice in the tacit game. The actual solutions might be different because the game contexts might be different, with different suggestive details; but the nature of the two solutions seems virtually identical since both depend on an agreement that is reached by tacit consent. This is true because the explicit agreement that is reached in the full-communication game corresponds to a priori expectations that were reached (or in theory could have been reached) jointly but independently by the two players before the bargaining started. And it is a tacit "agreement" in the sense that both can hold confident rational expectations only if both are aware that both accept the indicated solution in advance as the outcome that they both know they both expect. [Any serious (non-ritualistic) attempt to get more than this solution could almost be described as "wilful breach of contract" if it occurs on the part of a rational player with rational partner.]

There is a qualification to this point. With full information about each other's value systems and a homogenous set of gains to be divided, there may be an infinity of equivalent solutions, all yielding the same values to the two players, but no difficulty in agreeing on an arbitrary choice among this indifferent set. But tacit bargaining often requires a further degree of coordination, namely, a coordinated choice even among equivalent divisions of the gains. Negotiation over a boundary line in

In these terms the first (explicit) part of the Harsanyi hypothesis might be rephrased: that there is, in any bargaining-game situation (with perfect information about utilities), a particular outcome such that a rational player on either side can recognize that any rational player on either side would recognize it as the indicated "solution". The second (implicit) part of the hypothesis is that the particular outcome so recognized is determined by mathematical symmetry. The first we might call the "rational-solution" postulate; it is the second that constitutes the "symmetry" postulate.

The question to be explored, now, is whether the symmetry postulate is derived from the players' rationality -- the rationality of their expectations -- or must rest on other (perhaps empirical) grounds. Additionally, if it rests on other grounds, what are they and how firm is the support?

To pursue the first question, whether symmetry can be deduced from the rationality of the players' expectations, we can consider the rationality of the two players jointly, and inquire whether a jointly expected non-symmetrical outcome contradicts the rationality postulate. If two players confidently believe they share, and do share, the expectation of a particular outcome, and that outcome is not symmetrical in a mathematical sense, can we demonstrate that their expectations were irrational, hence that the rationality postulate is contradicted? Specifically, suppose that two

---

homogenous territory is thus different from the simultaneous dispatch of troops to take up positions representing "claims"; such claims may overlap and cause trouble even though the terrain values claimed are consistent. Thus the coordination problem is different; and there is no a priori assurance that the solution to the tacit game (or to games with somewhat incomplete communication, information, etc.) would be in the set of equivalent solutions to the fully explicit game.

players may have $100 to divide as soon as they agree explicitly on how to divide it; and they quite readily agree that A shall have $80 and B shall have $20; and we know that dollar amounts in this particular case are proportionate to utilities, and the players do too: can we demonstrate that the players have been irrational?

We must be careful not to make symmetry part of the definition of rationality; to do so would destroy the empirical relevance of the theory, and simply make symmetry an independent axiom. We must have a plausible definition of rationality that does not mention symmetry, and show that asymmetry in the bargaining expectations would be inconsistent with that definition. For our present purpose we must suppose that two players have picked $80 and $20 by agreement, and see whether we can identify any kind of intellectual error, misguided expectations, or disorderly self-interest, on the part of one or both of them, in their failure to pick a symmetrical point.

Specifically, where is the "error" in B's concession of $80 to A? He expected -- he may tell us, and suppose that we have means to check his veracity (a modest supposition if full information of utilities is already assumed!) -- that A would "demand" $80; he expected A to expect to get $80; he knew that A knew that he, B, expected to yield $80 and be content with $20; he knew that A knew that he knew this; and so on. A expected to get $80, knew that B was psychologically ready because he, B, knew that A confidently expected B to be ready, and so on. That is, they both knew -- they tell us -- and both knew that both knew, that the outcome would ineluctably be $80 for A and $20 for B. Both were correct in every expectation; the expectations of each were internally consistent

and consistent with the other's. We may be mystified about how they reached such expectations; but the feat claims admiration as much as contempt. The "rational-solution" postulate is beautifully borne out; the game seems to have "dictated" a particular outcome that both players confidently perceived. If, at this point, we feel we ourselves wouldn't have perceived the same outcome, we can conclude that one of four hypotheses is false: (1) the rational-solution postulate, (2) the rationality of A and B, (3) our own rationality, (4) the identity (in all essential respects) of the game that we introspectively play with the game that A and B have just played. But we cannot, on the evidence, declare the second to be the false one -- the rationality of A and B.

Note that if B had insisted on $50, or if A had been content to demand $50, claiming to be rational and arguing in terms of confidence in a shared expectation of that outcome, both players would have been in "error" and we cannot tell, on the evidence, which one is irrational or whether they both are. Unless we make symmetry the definition of rationality we can only conclude that at least one of the players is irrational or that the rational-solution postulate does not hold. What we have is a single necessary condition for the rationality of both players jointly; we have no sufficient condition, and no necessary condition that can be applied to a single player.

Nor can we catch them up if we ask them how they arrived at their expectations. Any grounds that are consistent would do, since any grounds that each expects the other confidently to adopt are grounds that he cannot rationally eschew. Consistent stories are all they need; and if they say that a sign on the blackboard said A-$80, B-$20, or that they saw in a

bulletin that two other players, named A' and B', split $80-$20, and that they confidently perceived that this was clear indication to both of them of what to expect, that this was the only "expectable" outcome, we cannot catch them in error and prove them irrational. They may be irrational; but the evidence will not show it.

There is, however, a basis for denying my present argument. Since I have not actually applied an independent test of rationality to two players, given them the game to play, and observed the 80/20 split that I just mentioned, but have only posed it as a possibility to see whether it would imply irrationality if it occurred, one could raise the objection that it could not possibly occur. And the argument would rest on the problem of coordination; it would run as follows.

If two players jointly expect a priori the same outcome, and confidently recognize it as their common expectation, they must have the intellectual power to pick a particular point in common. If the whole $100 can be divided to the nearest penny, there are 9,999 relevant divisions to consider, one of which would have to be picked simultaneously but separately by both players as their expectations of the outcome. But how can two people concert their selections of one item out of 9,999, in the sense that their expectations focus or converge on it, except with odds of 9,999 to 1 against them? The answer must be that they utilize some trick, or clue, or coordinating device that presents itself to them. They must, consciously or unconsciously, use a selection procedure that leads to unique results. There must be something about the point they pick that distinguishes it -- if not in their conscious reasoning, at least in our conscious analysis -- from the continuum of all possible alternatives.

Now, is it possible for two rational players, through anything other than sheer coincidence or magic, to focus their attention on the same particular outcome and "rationally" be confident that the other is focussed on the same outcome with the same appreciation that it is mutually expected? And, if so, how can they?

The answer is that they can; this was demonstrated by the fact that in a sample of over 40 people, instructed to concert tacitly on picking the same positive number from among the infinity of positive numbers, 40% picked the same number; and that, instructed to concert tacitly on picking a single cell from a 3 x 3 matrix or a 4 x 4 matrix, a substantial majority managed to do so in spite of enormous odds against them in a random sense. How can they? By using any means that is available: any clue, any suggestion, any rule of elimination, that leads to an unambiguous choice or a high probability of concerted choice. And one of these rules, or clues, or suggestions, is mathematical symmetry.*

In a game that has absolutely no details but its mathematical structure, in which no inadvertent contextual matter can make itself appreciated by a player as something that the other can appreciate too, there may be

---

* The basic intellectual premise, or working hypothesis, for rational players in this game seems to be the premise that some rule must be used if success is to exceed coincidence, and that the best rule to be found, whatever its rationalization, is consequently a rational rule. (This premise would support, for example, J. F. Nash's model that views an "unsmoothed" tacit game as the limit of a "smoothed" game as the smoothing approaches zero. While this view of the unsmoothed game is in no sense logically necessary it is a powerfully suggestive one that can, in the absence of any better rationale for converging on a single point, command the attention of players in need of a common choice. The limiting process provides a clue for picking one of the infinitely many equilibrium points that actually exist in the unsmoothed game. Of course, the premise equally supports any other procedure that produces a candidate for election among the infinitely many potential choices.)

nothing to work on but a continuum of numbers. And all the numbers can be sorted according to whether they correspond to symmetrical or asymmetrical divisions. If all numbers but one represent an asymmetrical split, then sheer mathematical symmetry is a sufficient rule and a supremely helpful one, in concerting on a common choice. And it may be possible to set up a game in such sanitary fashion, with suppression of identity of players and everything else, that there is literally no other visible basis for concerting unless impurities creep in.*

In other words, mathematical symmetry may focus the expectations of two rational players because it does -- granted the other assumed features of the game, like full information on each other's utility systems -- provide one means of concerting expectations. Whether it is a potent means may depend on what alternatives are available.

That there are other means of concerting, including some that may substantially outweigh the notion of symmetry, seems amply demonstrated

---

* In this view, the theory of Nash (leading to the maximum-utility-product solution) is a response to the fact that even in the realm of mathematics there are offhand too many types of uniqueness or symmetry to provide an unambiguous rule for selection, hence a need to adduce plausible criteria (axioms) sufficient to yield an unambiguous selection. Braithwaite's theory can be characterized the same way. The fact that the two solutions conflict implies that mathematicians may not have a sufficiently common mathematical esthetics to satisfy the first part of the Harsanyi postulate, i.e., to coordinate their expectations on the same outcome. (R.B. Braithwaite, Theory of Games as a Tool for Moral Philosophy, Cambridge University Press, 1955; Braithwaite's solution is described in Luce and Raiffa, Games and Decisions, New York: 1957, p. 145 ff.)

Braithwaite's construction of the problem as a one-person arbitration problem, and Luce and Raiffa's reformulation of Nash's theory in terms of arbitration rather than strategy (pp. 121-154), seem to emphasize that intellectual coordination is at the heart of the theory. A legalistic solution requires some rationalization of a unique outcome; pure casuistry is helpful if the alternative is vacuum.

by some of the experiments reported in the article already referred to.* So it is demonstrably possible to set up games in which mathematical symmetry does provide the focus for coordinated expectation, and demonstrably possible to set up games in which some other aspect of the game focusses expectations. (These other aspects are commonly not contained in the mathematical structure of the game but are part of the "topical content"; i.e., they usually depend on the "labelling" of players and strategies, to use the term of Luce and Raiffa.)

I have no basis for arguing with what force, or in what percentage of interesting games, mathematical symmetry does dominate "rational expectations". But I think that the status of the symmetry postulate is qualitatively changed by the admission that symmetry has competitors in the role of focussing expectations. For if it were believed that rational players' expectations could only be brought into consistency by some mathematical property of the payoff function, then symmetry might seem to have undisputed claim, particularly if it is possible to find a unique definition of symmetry that meets certain attractive axioms. But if one has to admit that other things, things not necessarily part of the mathematical structure of the payoff function, can do what symmetry does, then there is no <u>a priori</u> reason to suppose that what symmetry does is 99% or 1% of the job. The appeal of symmetry is no longer mathematical, it is introspective; and further argument is limited to the personal appeal of particular focussing devices to the game theorist <u>as game player</u>, or else recourse must be had to empirical observation.

---

* T. C. Schelling, "Bargaining, Communication, and Limited War," <u>Journal of Conflict Resolution</u>, Vol. 1, March, 1957.

Thus a normative theory of games, a theory of strategy, depending on intellectual coordination, has a component that is inherently empirical; it depends on how people can coordinate their expectations. It depends therefore on skill and on context. The rational player must address himself to the empirical question of how, in the particular context of his own game, two rational players might achieve tacit coordination of choices, if he is to find in the game a basis for sharing an a priori expectation of the outcome with his partner. The identification of symmetry with rationality rests on the assumption that there are certain intellectual processes that rational players are incapable of, namely, concerting choices on the basis of anything other than mathematical symmetry, and that rational players should know this; it is an empirical question whether rational players can actually do better than such a theory predicts, and should consequently ignore the strategic principles produced by the theory.

Final Note

An introspective game, which could be submitted to experiment, may illustrate the point. Let us -- whether or not we are strongly attracted to the symmetry postulate, and whether or not we are especially attracted to the particular symmetry of the Nash solution -- put ourselves in a frame of mind congenial to accepting the "Nash point" as the rational outcome of an explicit bargaining game. And imagine a game's potential payoffs as consisting of all the points on or within some boundary in the upper-right quadrant. We now consider some variants of this game.

First, we are to play the same game in tacit form; each of us picks a point along his own axis, and if the resulting point is on or within the

boundary, we get the amounts (utilities) denoted by the coordinates we pick. I conjecture that, in the frame of mind I have asked for, we should probably pick the Nash point. Without asking precisely why, let us go on to the next variant. This game is tacit too, but it differs in that we get nothing unless the point whose coordinates we pick is exactly on the boundary. We get nothing unless we exhaust the available gains. Caution gets us nowhere; each must choose exactly as the other expects him to. Again I propose that, in our present frame of mind, we ought to take the Nash point.

Finally consider another variant: we are shown the diagram of the game that has just been played and told that we are now to be perfect partners, winning and losing together. Conscious of the fact that our present game is modelled on a "bargaining game" we are to pick, without communicating, coordinates of a point that lies exactly on the boundary; if we do, we both win prizes -- the same prizes no matter what point we succeed in picking together -- and if we fail to pick a point on the boundary we get nothing. In this pure coordination game, I conjecture again that we ought (would) in our present frame of mind pick the Nash point.

Why? Simply because we need some rationalization that leads to a unique point; and in the context the bargaining analogy provides it. Unless there is a sharp corner (which is then likely to be the Nash point anyway); or a simple "mid point" as when the boundary is a straight line or circular arc (which again coincides with the Nash point); or some especially suggestive form that seems to point towards a particular point; or unless there is an impurity (such as a dot on the boundary, from a printer's error, or a single point whose coordinates are whole numbers,

etc.); we may be led to search for a "unique" definition of symmetry to fall back on, Nash-type symmetry being as plausible as any I can think of, not as simple as some (like the intersection of a 45-degree line with the origin and others of that ilk) but less ambiguous on its own level of sophistication.

And if the Nash point appeals to us powerfully in the bargaining game, it must do so because we are confident that it appeals equally to our partner who in turn we believe to be aware that our views coincide; it must therefore appeal to us in the pure-coordination game as a unique point that the partner will consider to be obviously obvious.

What does this prove or suggest? I am not arguing for the Nash point. I am arguing rather that the appeal of the Nash point to a game theorist (as introspective game player) may be the reverse of the sequence I have just run through. It may be the focal quality of the Nash-point in the pure coordination game -- the unequivocal usefulness of a uniquely defined symmetry concept, when no non-mathematical impurities are available to help -- that makes it a controlling influence in the tacit and terribly cooperative boundary-line game; that in turn makes it a reliable guide in the less demanding tacit bounded-area game; and that in turn takes the heart out of any player in the explicit game who thinks expectations can focus anywhere else.

In other words, by postulating the need for coordination of expectations, we seem to have a theoretical basis for something like the Nash axioms. What a theory like Nash's needs is the premise that a "solution" exists; it is the observable phenomenon of tacit coordination that provides empirical evidence that (sometimes) rational expectations can be tacitly

focussed on a unique (and perhaps efficient) outcome, and that leads one to suppose that the same may be possible in a game that provides nothing but mathematical properties to work on. The Nash theory is vindication of this supposition -- complete vindication if it dominates all competing mathematical solutions in terms of mathematical esthetics. (The resulting focal point is limited to the universe of mathematics, however, which should not be equated with the universe of game theory.)